Biometria: Reconhecimento de voz

Sales, Stanley Andreato
Clínica de Nutrição Ygiís e Instituição de Ensino – ME
stanleysales@yahoo.com

Sales, Nevilde M. R.
Clínica de Nutrição Ygiís e Instituição de Ensino – ME
nevilderiselo@yahoo.com.br

Silva, Marco Antônio L. R.
Fundação Getúlio Vargas
profmarcoantonio@gmail.com

RESUMO

O paradigma mundial da inclusão social se tornou, atualmente, um dos maiores desafios enfrentados pelos governantes e sociedade civil. Para indivíduos portadores de necessidades especiais, a busca por assegurar condições que os integrem a comunidade é de extrema importância. Assim, o reconhecimento de voz se tornou um aliado de grande relevância pois, através da fala, o indivíduo conseguirá efetuar alguns comandos como ligar a televisão, mudar de canal e/ou digitar/ler em um computador, tablet ou celular o que se torna quase impossível sem a ajuda de outrem. **Objetivo:** demonstrar como desenvolver um protótipo simples capaz de reconhecer a voz humana e interpretá-la como comando para um computador através da comparação da palavra pronunciada com amostras pré-cadastradas. **Metodologia:** implementação de um protótipo de reconhecimento de voz desenvolvido em Microsoft Visual Basic. **Conclusão:** O protótipo desenvolvido obteve resultados compatíveis com os existentes na comparação através da correlação entre o sinal de entrada e amostras pré-cadastradas se mostrando plausível a futuros estudos e aprimoramento para que possa ser utilizado comercialmente.

Palavras-chaves

Biometria, processamento de sinal digital, reconhecimento de voz.

ABSTRACT

The global paradigm of social inclusion has become, currently, one of the biggest challenges faced by government and civil society. In individuals with special needs, the search for ensuring conditions that integrate them into the community is extremely important. Thus, speech recognition has become an ally of great relevance, because through speech, the individual will be able to carry out some commands such as turning on the television, changing channels and / or typing / reading on a computer, tablet, cell phone becomes almost impossible without the help of others. **Objective**: to demonstrate how to develop a simple prototype capable of recognizing the human voice and interpreting it as a command for a computer by comparing the pronounced word with pre-registered samples. **Methodology**: implementation of a speech recognition prototype developed in Microsoft Visual Basic. **Conclusion**: The developed prototype obtained results compatible with those existing in the comparison through the correlation between the input signal and pre-registered samples.

Keywords

Biometry, digital signal processing. Voice recognition

1. INTRODUÇÃO

O paradigma mundial da inclusão social, se tornou, atualmente, um dos maiores desafios enfrentados pelos governantes e sociedade civil. No caso de indivíduos portadores de necessidades especiais, a busca por assegurar condições que integrem esses indivíduos a comunidade é de extrema relevância [1]. A acessibilidade de indivíduos cegos/com baixa visão ou com restrição de movimentos torna essa inclusão social um desafio pois, para estes indivíduos, tarefas simples como ligar a televisão, mudar de canal e/ou digitar/ler em um computador, *tablet,* celular se torna quase impossível sem a ajuda de outrem.

Atualmente, vários equipamentos eletrônicos empregam algum tipo de biometria para realizar tarefas que vão de simples a complexas. Como exemplo, há máquinas fotográficas que possuem um sistema de reconhecimento de face e sorriso, brinquedos que reagem a comandos de voz ou quaisquer outros sons, alarmes que disparam quando detectam ruídos, câmeras que detectam movimento para então iniciarem a gravação de uma cena. Há, ainda, computadores que podem escrever textos ou executar tarefas através de comandos de voz ou através da detecção de movimento do globo ocular. Existem equipamentos instalados em veículos capazes de calcular o tempo de fechamento do olho do motorista pela velocidade de movimento da pálpebra, inferindo se o motorista está sonolento ao volante ou não e emitindo um aviso em caso afirmativo.

Tarefas de maior ou menor relevância como segurança/saúde e lazer/entretenimento respectivamente, podem ser executadas por sistemas computacionais através do emprego da biometria, porém estes recursos não estão disponíveis a todos por, normalmente, terem um custo agregado.

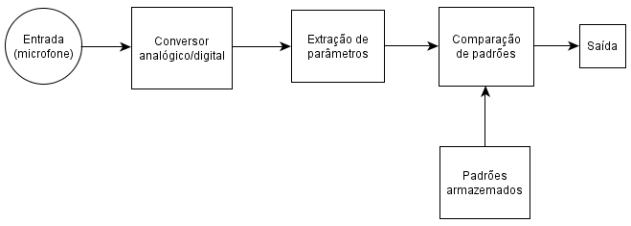
A biometria consiste de métodos para reconhecer uma pessoa, baseada em suas características comportamentais e/ou fisiológicas únicas [2, 3]. Essas características incluem impressão digital, voz, face, retina, íris, assinatura, geometria das mãos, veias dos pulsos, dentre outras. De todas as características mencionadas, a que os humanos aprendem a reconhecer primeiramente é a fala, ou características da voz [4, 5].
A fala é um modo natural de comunicação para as pessoas. Adquirimos a habilidade de falar durante a infância e

continuamos a depender da fala por toda a vida. É tão natural que não percebemos o quão complexo é o fenômeno da fala. O trato vocal humano e articuladores são órgãos biológicos com propriedades não-lineares, cuja operação não está sob controle consciente e também pode ser afetada por fatores que vão desde o gênero/idade até fatores emocionais [4, 5]. Como resultado, vocalizações podem variar muito em termos de sotaque, pronúncia, articulação, nasalidade, tom, volume e velocidade; ainda, durante a transmissão, os padrões irregulares da fala podem ser distorcidos por ruídos de fundo e ecos, assim como por características elétricas (no caso de a comunicação estar sendo efetuada através de uma linha telefônica, por exemplo) [4, 5].

Devido à capacidade do nosso cérebro de interpretar informações extremamente complexas, podemos, de forma praticamente inconsciente, captar em uma mensagem falada várias informações além da transmitida textualmente pelas frases vocalizadas. Reconhecemos assim quem nos está falando, sua posição no espaço físico, seu estado emocional e vários outros dados que podem estar escondidos no tom de voz usado (ironia, seriedade ou tristeza, por exemplo) [4, 5]. O reconhecimento de voz, segundo Foster [6] é o processo de identificação automática de palavras faladas. Para Tebelskis [7] é uma tarefa de reconhecimento de padrões multinível, na qual sinais acústicos são examinados e estruturados em uma hierarquia de unidades de subpalavras (fonemas), palavras, frases e sentenças.

Desde os anos 1970, na área de reconhecimento de voz grandes progressos têm ocorrido. Atualmente se utiliza uma série de abordagens de engenharia como comparação de padrões, modelagem estatística [2], explica que os sinais da fala são gravados em arquivos de áudio em um computador, utilizando-se programas apropriados para essa tarefa. As informações nesses arquivos são convertidas do domínio do tempo para o domínio da frequência utilizando-se de técnicas de processamento de sinal digital. O espectro de frequências é então utilizado para treinar o sistema de reconhecimento de voz, conforme demonstrado no diagrama abaixo, **Figura 1**:

Figura 1: Diagrama de blocos de um sistema típico de reconhecimento de voz.



Fonte: Sales et al. 2009

O sistema reconhece a voz do locutor comparando o sinal da entrada com os padrões armazenados na memória do computador. Quando um dos padrões coincide, a saída é acionada. Para isso, uma análise do sinal é efetuada e são extraídas informações úteis do mesmo a serem processadas posteriormente, como por exemplo, na comparação de padrões. Segundo Tebelskis [7] e Walker [8], diversas técnicas de análise de sinal podem ser empregadas, dentre elas a de *Fast Fourier Transform* (FFT).

Entretanto, os computadores ainda estão longe do nível de desempenho humano no reconhecimento de voz. Além da arquitetura dos computadores, fatores externos contribuem para a ineficiência dos reconhecedores de voz (sistemas de reconhecimento de voz).

Foster [6] explica que reconhecedores de voz disponíveis comercialmente abrangem uma vasta área de desempenho. Para avaliar o desempenho, é preciso definir e medir a precisão do reconhecimento. Essa precisão depende do tipo de sistema (se dependente ou independente do locutor), do tipo de reconhecimento de palavra (se discreto ou contínuo), da dificuldade do vocabulário, da qualidade da tecnologia e do ambiente do usuário.

Apesar de não haver um sistema 100% eficiente, os reconhecimentos de voz por computadores, bem como outras características da biometria, estão sendo cada vez mais difundidos no cotidiano das pessoas, para sua comodidade, diversão, segurança e inclusão social de portadores de necessidades especiais sendo de grande relevância estudos cada vez mais profundos sobre essa tecnologia.

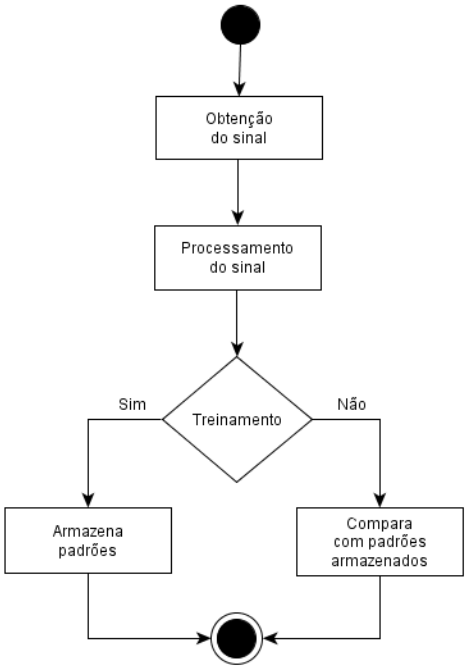
Assim, o objetivo deste trabalho (protótipo) é a implementação de um aplicativo simples, de fácil execução, baixo custo, sem utilização de produtos consolidados de terceiros e que seja capaz de reconhecer a voz humana e interpretá-la como comando ao computador através da comparação da palavra pronunciada com amostras pré-cadastradas. Esse estudo teve como premissa ser uma fonte de conhecimento/pesquisa da técnica de construção de software para o reconhecimento de voz através de linguagem de programação simples, de fácil execução para a arquitetura de programas. Assim, o desafio deste protótipo desenvolvido foi a implementação de um aplicativo capaz de reconhecer a voz humana e interpretá-la sob forma de comando.

2. METODOLOGIA

O programa foi escrito no Microsoft Visual Basic versão 6. Todos os algoritmos para o reconhecimento da voz foram escritos nessa linguagem sem o emprego de bibliotecas de terceiros. Foi escolhida a linguagem Microsoft Visual Basic versão 6 por ser de fácil entendimento e de desenvolvimento, para demonstrar técnicas utilizadas em programas de reconhecimento de voz de forma simples e compreensível (o código fonte não será disponibilizado por ter direitos autorais).

Para a implementação do protótipo foi desenvolvido um diagrama simplificado de fluxo conforme **Figura 2**

Figura 2: Diagrama simplificado de fluxo do protótipo



Fonte: Sales et al, 2009

No protótipo, o aplicativo captura os sinais de entrada do microfone pelo tempo de um segundo. Após decorrido esse tempo, desliga o microfone e processa o sinal recebido. Se a função de reconhecimento estiver ativada, compara o padrão do sinal recebido com um padrão previamente armazenado e retorna sua correlação.

Para acionar a placa de som e iniciar a captura de sinal pelo microfone, o aplicativo executa a interface de programação de aplicativos *Application Programming Interface* (API) *mciSendStringA* da biblioteca de ligação dinâmica *Dynamic-link library* (DLL) WINMM.DLL. Essa API somente habilita o programa a capturar o áudio da placa de som. Todo o restante do processamento referente ao reconhecimento da voz foi realizado através de algoritmos próprios utilizando o cálculo da *Fast Fourier Transform* (FFT), [8].

Para determinar as frequências componentes do sinal, utilizou-se a Transformada Discreta de Fourier (DFT). A DFT de comprimento N na sequência x_k correspondente ao sinal é definida por:

$$x_k = \sum_{j=0}^{N-1} X_j e^{-i2\pi jk/N}$$

para todos os inteiros = 0, ±1, ±2,...

Conforme Walker [8], a DFT transforma a sequência x_k do domínio do tempo para o domínio da frequência, tornando possível visualizar as frequências componentes desse sinal e assim extrair a informação desejada com um menor número de dados em relação ao sinal complexo.

Um melhoramento da DFT, chamado Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*), foi desenvolvido e, este, reduz o tempo de cálculo pelo fator de duzentos quando N = 1024. A FFT é definida por:

$$x_k = \sum_{j=0}^{N-1} X_j W^{jk}$$

Onde W pode ser $e^{-i2\pi/N}$ ou $e^{i2\pi/N}$.

O protótipo emprega a FFT como técnica de extração de informação do sinal de voz a ser comparado com um padrão previamente armazenado. Uma vez que a FFT tem como resultado uma sequência de números complexos $X[k]$ mesmo quando $x[n]$ são reais, essa comparação é feita através da correlação entre a sequência $X[k]$ e a média das sequências $X[k]$ armazenadas.

A correlação dos padrões foi implementada a partir do coeficiente de correlação de Pearson, [9, 10], que é expresso pela fórmula:

$$r = \frac{S_{xy}}{S_x S_y}$$

Onde, S_{xy} = covariância entre as duas variáveis, obtida através da fórmula:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

S_x e S_y = desvios-padrão das variáveis X e Y. Sendo expressos por:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \quad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y - \overline{y})^2}{n-1}}$$ e

Resultando na fórmula do coeficiente de correlação de Pearson:

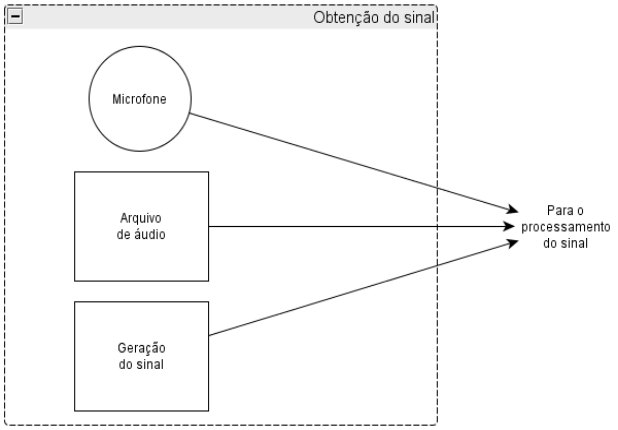
$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Onde, n = total de itens da sequência
r = coeficiente de correlação
$\overline{x}$ = média aritmética da sequência X
$\overline{y}$ = média aritmética da sequência Y

Conforme Barbetta [9] e Bisquerra [10], a correlação será tanto maior quanto mais distante de zero estiver r até os limites máximos de 1 ou -1 ($-1 \leq r \leq 1$), ou seja, a correlação é perfeita se r = 1 ou r = -1.

A obtenção do sinal deu-se de três formas: através do microfone, de um arquivo de áudio ou pela geração de sinal no próprio aplicativo, conforme diagrama abaixo (**Figura 3**):

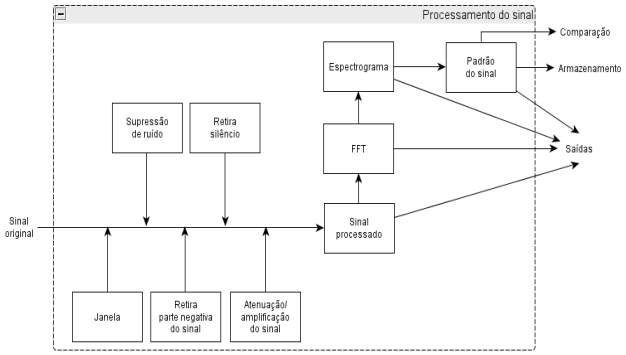
Figura 3: Diagrama de blocos mostrando as três formas de obtenção de sinal.



Fonte: Sales et al, 2009

O sinal de entrada foi normalizado [11] para que variações entre o que foi armazenado como padrão e o que está sendo dito para o reconhecimento tivessem o mínimo de interferência no processo de reconhecimento, conforme descrito na **Figura 4**.

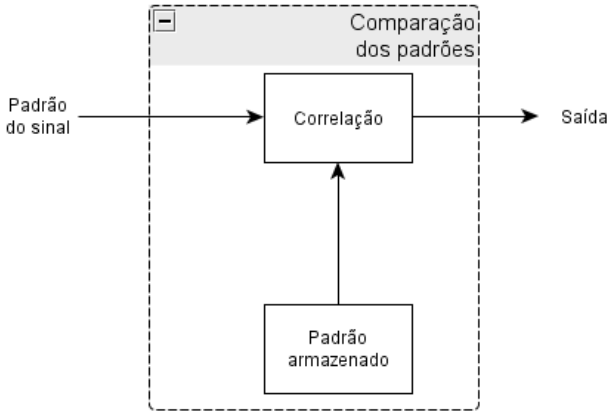
Figura 4: Diagrama de blocos da etapa de processamento do sinal.



Fonte: Sales et al, 2009

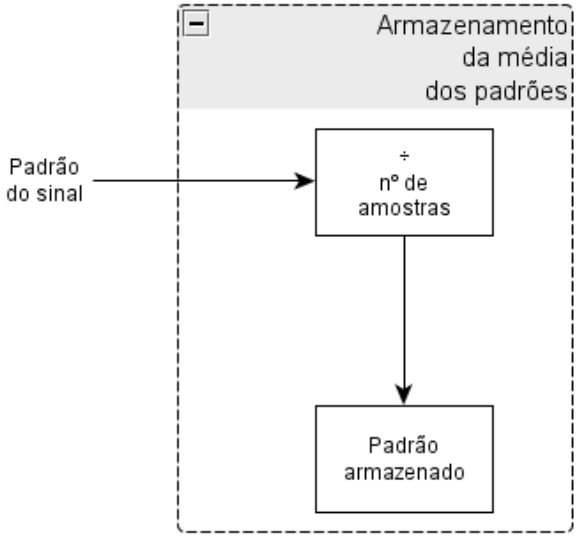
A comparação do sinal foi realizada verificando-se a correlação entre o padrão calculado e o padrão armazenado. O padrão calculado é a média dos espectrogramas decompostos em quadrantes e o padrão armazenado está em um banco de dados e corresponde à média dos padrões calculados em cada amostra. No protótipo, o usuário determina o número de amostras por palavra, o que significa que a palavra deve ser repetida a quantidade especificada no número de amostras. Cada vez que a palavra é falada sofre variações naturais, acarretando em vários padrões gerados. A média desses padrões é então obtida e armazenada para futura comparação, conforme **Figura 5 e Figura 6** abaixo.

Figura 5: Diagrama de blocos da comparação entre os padrões



Fonte: Sales et al, 2009

Figura 6: Diagrama de blocos representando a forma de armazenamento dos padrões amostrados.



Fonte: Sales et al, 2009

2.1 **Testes**

Existem, atualmente, muitos softwares comerciais de reconhecimento de voz, por exemplo: *SpeechExec* e *FreeSpeech* da *Philips*, *Dragon NaturallySpeaking* da *Dragon/Nuance Systems*, *MacSpeech Dictate* da *MacSpeech*, *TalkIt TypeIt 2 Deluxe/Easy Letters* da *PC Treasures*, *iTranslate da Ecato.etc.*.
Para os testes de comparação de reconhecimento de voz foram utilizados como referência o programa *ViaVoice* da IBM por ser reconhecido internacionalmente.

Como o aplicativo reconhece somente palavras isoladas, as palavras das frases foram gravadas uma a uma no seu banco de dados. Foi utilizada a seguinte configuração do aplicativo para a geração dos padrões:

- Supressão de 10% do ruído ambiente;
- Não houve atenuação/amplificação do sinal de entrada;
- Foi retirado o silêncio inicial do sinal, ou seja, a parte do sinal entre o início da captura e o início da fala;
- Foram retirados os valores correspondentes ao silêncio em todo o sinal (valor zero);
- Foram utilizados 256 pontos para a FFT, sem sobreposição;
- Não foi empregada qualquer janela no sinal de entrada;
- O tamanho do padrão foi de 15x10 pois com esse tamanho foi obtida a melhor relação entre o desempenho de processamento e reconhecimento de voz, ou seja, o espectrograma foi mapeado em 15 linhas por 10 colunas, gerando um total de 150 quadrantes;

O sinal de entrada foi capturado à taxa de 44,1KHz, estéreo (mas apenas um canal foi utilizado para processamento) e 16 bits de precisão. O nível do microfone estava a 50% e 0dB de ganho no sistema operacional.

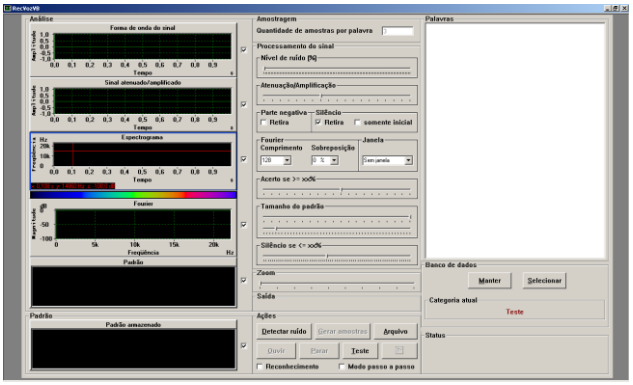
O treinamento do aplicativo se deu com a geração de padrões das palavras para locutor masculino, feminino e masculino/feminino. Foram criadas amostras em quintuplicatas para cada palavra criando um conjunto de padrões com cinco amostras para o locutor masculino e cinco amostras para o locutor feminino.

Foram geradas cinco amostras por palavra (cinco amostras para cada locutor) com isso foi possível criar um conjunto de padrões com amostras para o locutor masculino e locutor feminino, formando um padrão misto. Porém, este teste não foi comparado com o programa de referência pois, o software *ViaVoice* não oferece opção de padrão misto.

2.2 **Resultados**

A apresentação gráfica do protótipo dividiu-se em três partes: 1) A coluna mais à esquerda **(Figura 7**) destina-se à apresentação gráfica do sinal, seja em sua forma de onda e/ou espectrograma. Apresenta ainda o gráfico da FFT calculada sobre o sinal, o padrão desse sinal e o padrão dos sinais armazenados para comparação. Até que todos esses gráficos sejam calculados e exibidos, o tempo entre a captura do sinal e o reconhecimento de seu conteúdo é elevado. O aplicativo captura apenas um segundo de sinal por vez, ele não reconhece palavras conectadas e discurso contínuo, ou seja, é um reconhecedor de palavras isoladas (discretas) somente. 2) A coluna central (**Figura 7**) contém os controles de processamento do sinal, bem como controles de visualização e botões de ações do usuário. 3) A coluna mais à direita **(Figura 7**) vai exibir a lista de palavras cadastradas e seu grau de coincidência quando do reconhecimento da voz. Há também, nesta coluna, os botões de manutenção e seleção do banco de dados que contém as palavras e padrões de reconhecimento.

Figura 7: Tela principal do aplicativo de reconhecimento de voz escrito em Visual Basic 6.



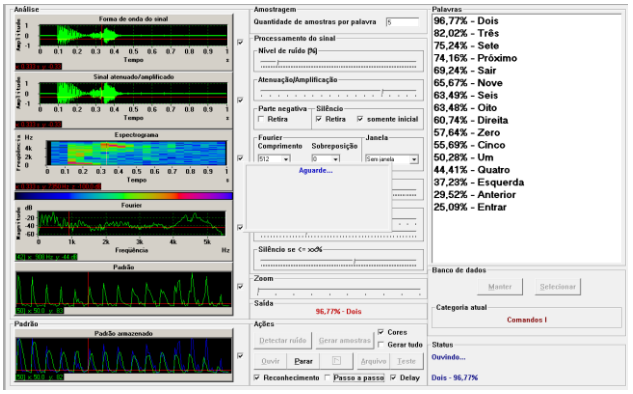
Fonte: Sales et al, 2009.

Foram criados padrões de fala utilizando dezesseis palavras e números aleatórios. Após criados estes padrões, o protótipo foi testado utilizando as mesmas palavras onde colocou-se o aplicativo no modo "Reconhecimento de Voz" e, em seguida, os locutores, masculino e feminino, falaram as palavras/números pré-definidas ao microfone. As **Figura 8** e **Figura 9** mostram a tela do aplicativo exibindo resultados da comparação entre as palavras/números falados e o padrão criado.

Nos resultados foi possível verificar acerto de 96,77% no numeral "dois" e 96,62% na palavra "sair". Em vermelho na área central das **Figura 8** e **Figura 9** é possível verificar o protótipo reconhecendo a palavra dita e mostrando o resultado.

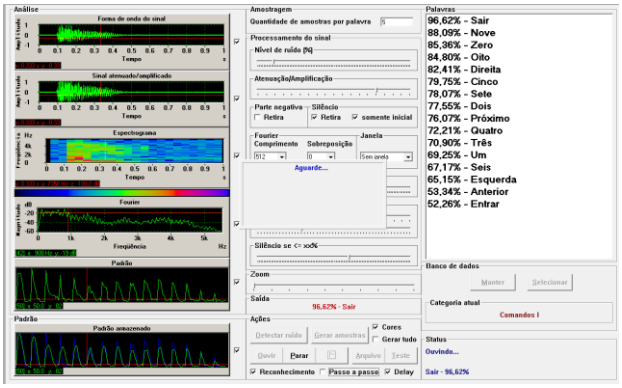
As demais palavras/numerais, quando testados, obtiveram os seguintes resultados: três (94,67%); próximo (95,55%); oito (94,08%); nove (93,29%); esquerda (91,88%); zero (91,77%) e anterior (91,67%) quando testadas também obtiveram um percentual de acerto superior a 90%. Já as palavras/numerais um (89,93%); cinco (89,56%); quatro (88,54%); sete (86,25%); entrar (87,17%); seis (81,79%) e direita (80,52%) obtiveram um percentual de acertos entre 80-90%.

Figura 8: Tela do aplicativo exibindo a comparação entre o padrão calculado de uma palavra falada e os padrões armazenados.



Fonte: Sales et al, 2009.

Figura 9: Tela do aplicativo exibindo a comparação entre o padrão calculado de uma palavra falada e os padrões armazenados.



Fonte: Sales et al, 2009.

Também foram testados frases e nome de letras do alfabeto aleatórias, utilizando locutores e padrões femininos, masculinos e misto. Nestes testes de reconhecimento de voz, no protótipo, foram obtidos os seguintes resultados, em porcentagem (

Tabela 1). Estes resultados se deveram ao fato de algumas letras (dicção das letras) serem foneticamente semelhantes, exemplos: B e P, F e S, dentre outras.

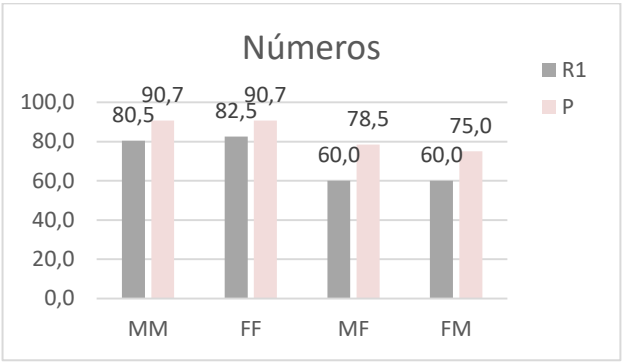
Tabela 1: Índices de acerto do protótipo de reconhecimento de voz desenvolvido em Visual Basic 6.

Locutor	Padrão gravado	Acerto de palavras em frases (%)	Acerto de nomes das letras (%)
Masculino	Masculino	81,38	60,00
Masculino	Feminino	41,72	30,00
Masculino	*Masculino/Feminino	62,07	40,00
Feminino	Masculino	40,69	30,00
Feminino	Feminino	84,42	40,00
Feminino	*Masculino/Feminino	64,48	30,00

Fonte: Sales et al., 2009. Obs.: * Teste misto (locutores masculinos e feminino juntos)

Quando testada a dicção utilizando numerais aleatórios, no protótipo, se obteve resultados estatisticamente significativos quando comparado com a referência conforme verificado na **Tabela 2.**

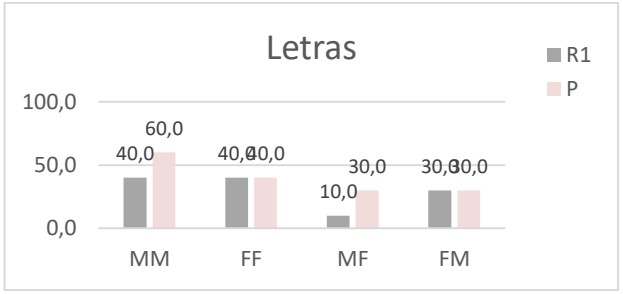
Tabela 2: Comparação dos resultados quando testado o reconhecimento de números.



Legenda - R1: *Via Voice*; P: Protótipo. MM: locutor masculino, padrão masculino; FF: locutor feminino, padrão feminino. MF: locutor masculino, padrão feminino; FM: locutor feminino, padrão masculino

Quando testada a dicção de letras aleatórias, no protótipo, foi obtido resultado estatisticamente significativo quando comparado com a referência conforme observado na **Tabela 3.**

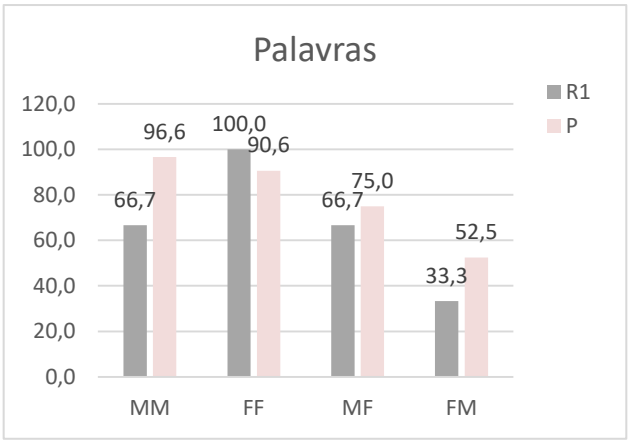
Tabela 3: Comparação dos resultados quando testado o reconhecimento de letras aleatórias do alfabeto.



Legenda - R1: *Via Voice*; P: Protótipo. MM: locutor masculino, padrão masculino; FF: locutor feminino, padrão feminino. MF: locutor masculino, padrão feminino; FM: locutor feminino, padrão masculino

O reconhecimento de voz, no protótipo, quando testado utilizando palavras foram obtidos resultados igualmente relevantes quando comparado com o programa de referência, conforme verificado na **Tabela 4.**

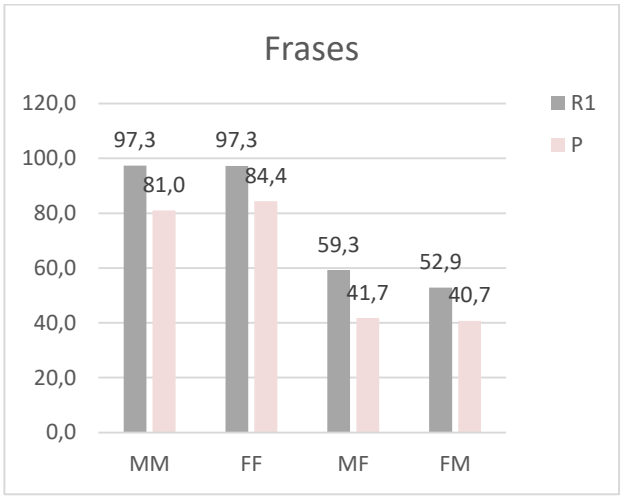
Tabela 4: Comparação do reconhecimento de voz quando utilizadas palavras isoladas.



Legenda - R1: *ViaVoice*; P: Protótipo. MM: locutor masculino, padrão masculino; FF: locutor feminino, padrão feminino. MF: locutor masculino, padrão feminino; FM: locutor feminino, padrão masculino

Nos testes com frases, porém, não foram obtidos resultados relevantes, conforme verificado na **Tabela 5** quando comparado com o programa de referência.

Tabela 5: comparação de resultados com software de referência



Legenda - R1: *ViaVoice*; P: Protótipo. MM: locutor masculino, padrão masculino; FF: locutor feminino, padrão feminino. MF: locutor masculino, padrão feminino; FM: locutor feminino, padrão masculino

2.3 **Discussão e conclusão**

Somos capazes de ouvir e entender uma pessoa falando mesmo entre várias outras conversando ao mesmo tempo. Subconscientemente filtramos as conversas e sons que não nos interessam. Essa habilidade de filtragem está além da capacidade dos sistemas de reconhecimento atuais. Reconhecimento da fala não é entendimento da fala. Entender o significado das palavras é uma função intelectual maior. Se um computador pode responder a um comando vocal, não significa que ele entende o comando falado.

Sistemas de reconhecimento da fala poderão, algum dia, ter a capacidade de distinguir nuances linguísticas e o significado das palavras. Por ora, os sistemas comerciais não atingem cem por cento de acerto todo o tempo, como foi possível observar nos testes realizados com o *ViaVoice* da IBM corroborando com os dados verificados, também, em nosso protótipo (

Tabela 1**; Tabela 3; Tabela 4; Tabela 5 eTabela 5**).

Foster [6] afirma que sistemas disponíveis comercialmente abrangem uma extensa gama de desempenho e esse desempenho é conseguido medindo-se a precisão do reconhecimento, que depende: a) do tipo de sistema (dependente ou independente de locutor); b) do tipo de reconhecedor (discreto, conectado ou contínuo); c) da complexidade do vocabulário; d) da qualidade da tecnologia empregada; e) do ambiente do usuário.

Para Foster [6] e Rabiner [3], um reconhecedor pode cometer três tipos de erros, sendo: 1) erro de substituição; 2) erro de rejeição; 3) erro de resposta espúria.

Foster [6] e Rabiner [3] classificam o erro de substituição como o mais crítico. Um erro de substituição ocorre quando o reconhecedor substitui a palavra dita por uma palavra incorreta. A taxa de erro de substituição deve ser menos de dois por cento para aceitação pelos usuários e, um erro de rejeição, ocorre quando o reconhecedor não classifica a palavra falada e a rejeita. Quando uma rejeição ocorre o usuário deve repetir a palavra até que o reconhecedor a identifique, erros de rejeição não são tão críticos como os de substituição e sua taxa deve ser inferior a três por cento.

Um erro de resposta espúria ocorre quando o reconhecedor classifica um som ou uma palavra inválida como uma palavra válida. Foster [6] afirma que nenhum reconhecedor está imune aos erros de resposta espúria e, quanto maior for o vocabulário, maiores as chances desse tipo de erro ocorrer. Foster [6] explica que tanto a taxa de erros de rejeição quanto a taxa de erro de substituição determinam a precisão do reconhecedor. Ele define que a taxa média de erro no reconhecimento é a soma da taxa de erro de substituição e a metade da taxa de erro de rejeição. Já Tebelskis [7] cita dois outros fatores dos quais depende a precisão de um sistema de reconhecimento de voz, sendo: 1). Restrições da linguagem (gramática) e 2). Discurso lido ou espontâneo (nesse último a fala pode conter gaguez, tosse, risada, palavras não terminadas, entre outras). Para ele o enfoque no reconhecimento da fala em processos passados se deu em três principais categorias: 1). Baseado em padrões – no qual uma palavra é comparada com padrões previamente gravados. Tem a vantagem de utilizar modelos de palavras precisos e a desvantagem de que os modelos são fixos, ou seja, variações da fala podem ser modeladas somente utilizando-se vários modelos por palavra; 2). Baseado em conhecimento – no

qual as variações da fala são codificadas no programa. A vantagem é que as variações são explicitamente modeladas e a desvantagem é que tal conhecimento é difícil de ser alcançado; 3). Baseado em estatística – no qual as variações da fala são modeladas estatisticamente, utilizando-se procedimentos de aprendizagem automáticos. Esse enfoque representa o atual estado da arte.

Diferentes distâncias do microfone em relação ao locutor e ruídos ambientes foram os principais fatores para a ocorrência de erros durante o processo de reconhecimento, tanto pelos *softwares* comerciais quanto pelo protótipo (

Tabela 1**; Tabela 3; Tabela 4; Tabela 5 e Tabela 5**). Palavras foneticamente semelhantes (homônimas e parônimas) também contribuíram para que o índice de acerto não chegasse a cem por cento, uma vez que essas palavras geraram erros de substituição nos aplicativos.

Com o protótipo de reconhecimento proposto neste estudo, as configurações estabelecidas e o conjunto de testes realizados, conseguiu-se uma média de acertos razoável para as palavras isoladas. Verificou-se que, dentre outras alterações na configuração do mesmo, um maior número de amostras por palavra foi fator relevante para o aumento da taxa de acertos durante o reconhecimento da voz. O aumento do número de pontos da FFT também contribuiu para uma melhor eficiência do protótipo, entretanto exigiu que as palavras faladas, durante o processo de reconhecimento, fossem ditas da forma mais semelhante possível à amostra armazenada, diminuindo a flexibilidade na pronúncia das mesmas. Uma maior eficiência juntamente com maior flexibilidade (liberdade) na pronúncia das palavras foi conseguida através da FFT com duzentos e cinquenta e seis ou quinhentos e doze pontos (

Tabela 1**; Tabela 3; Tabela 4; Tabela 5 e Tabela 5**).

O protótipo, assim como o *software* utilizado com referência de comparação, se mostrou dependente de locutor, uma vez que maiores taxas de acerto ocorreram quando os locutores, masculino e feminino, ditavam palavras para reconhecimento empregando seus respectivos padrões armazenados.

Estudos mais aprofundados sobre as estruturas morfológicas das palavras e processamento digital de sinal podem ser objetos de pesquisas futuras para o melhoramento do desempenho do protótipo, inclusive o emprego de redes neurais que possibilitarão, ao protótipo, aprender com os erros e evitá-los, aumentando consideravelmente seu desempenho e possibilitando seu uso comercial.

Os autores relatam não haver conflito de interesses, todos participaram igualmente na elaboração do projeto.

3. Referência Bibliográfica

[1] Fonseca, Ricardo T. M. 2005. *Os direitos Humanos e a pessoa com deficiência no mercado de trabalho*. In: Inclusão: Revista da Educação Especial. Brasília: Secretaria da Educação Especial/ MEC, 1,1, p. 19-24. (Out. 2005) http://portal.mec.gov.br/seesp/arquivos/pdf/revistainclusao2.pdf

[2]. Venayagamoorthy, G. K.; Moonasar, V.; Sandrasegaran, K. 1998. *Voice recognition and Signal processing*. COMSIG apos; 98. Proceedings of the 1998 South African Symposium on, 7-8, p. 29-32, (Set. 1998). **Doi=** 10.1109/COMSIG.1998.736916

[3] Rabiner, L. e Juang, B. H. 1993. *Fundamentals of speech recognition*. New Jersey: PTR Prentice-Hall. ISBN: 0-13-015157-2

[4] Guyton, Arthur C.; Hall, J. E. 2011. *Tratado de fisiologia médica*. ELSEVIER; Edição: 12,433 pg. ISBN- 978-85-352-6285-8

[5] Dangelo, J. G. 2007. Anatomia Humana Sistêmica e Segmentar. 3 ed., Atheneu, São Paulo, 763p. ISBN: 85-7379-848-3.

[6] Foster, P. R. e Schalk, T. B. 1993. *Speech recognition: the complete practical reference guide*. 1ª ed. New York: Flatiron Publishing. ISBN: 0-9366648-39-2

[7] Tebelskis, J. 1995. *Speech recognition using neural networks*. Thesis (Doctor of Philosophy in Computer Science) – School of Computer Science Carnegie Mellon University, Pittisburg, Pensylvania. https://isl.anthropomatik.kit.edu/pdf/Tebelskis1995.pdf

[8] Walker, J. S**.** *Fast Fourier transforms***.** 1996. 2ª ed. Florida: CRC Press. ISBN: 0-8493-7163-5

[9] Barbetta, P. A.; Reis, M. M.; Bornia, A. 2008. C. *Estatística para cursos de engenharia e informática*. 2ª ed. São Paulo: Atlas. ISBN: 978-85-224-5994-0.

[10] Bisquerra, R.; Sarriera, J. C.; Martinéz, F. 2004. *Introdução à estatística*. Porto Alegre: Artmed. ISBN: 978-85-363-0196-9.

[11] Mitra, S. K. 2005. *Digital signal processing: a computer based approach*. 3ª ed. New York: McGraw-Hill. ISBN: 978-0-07-286546-2.